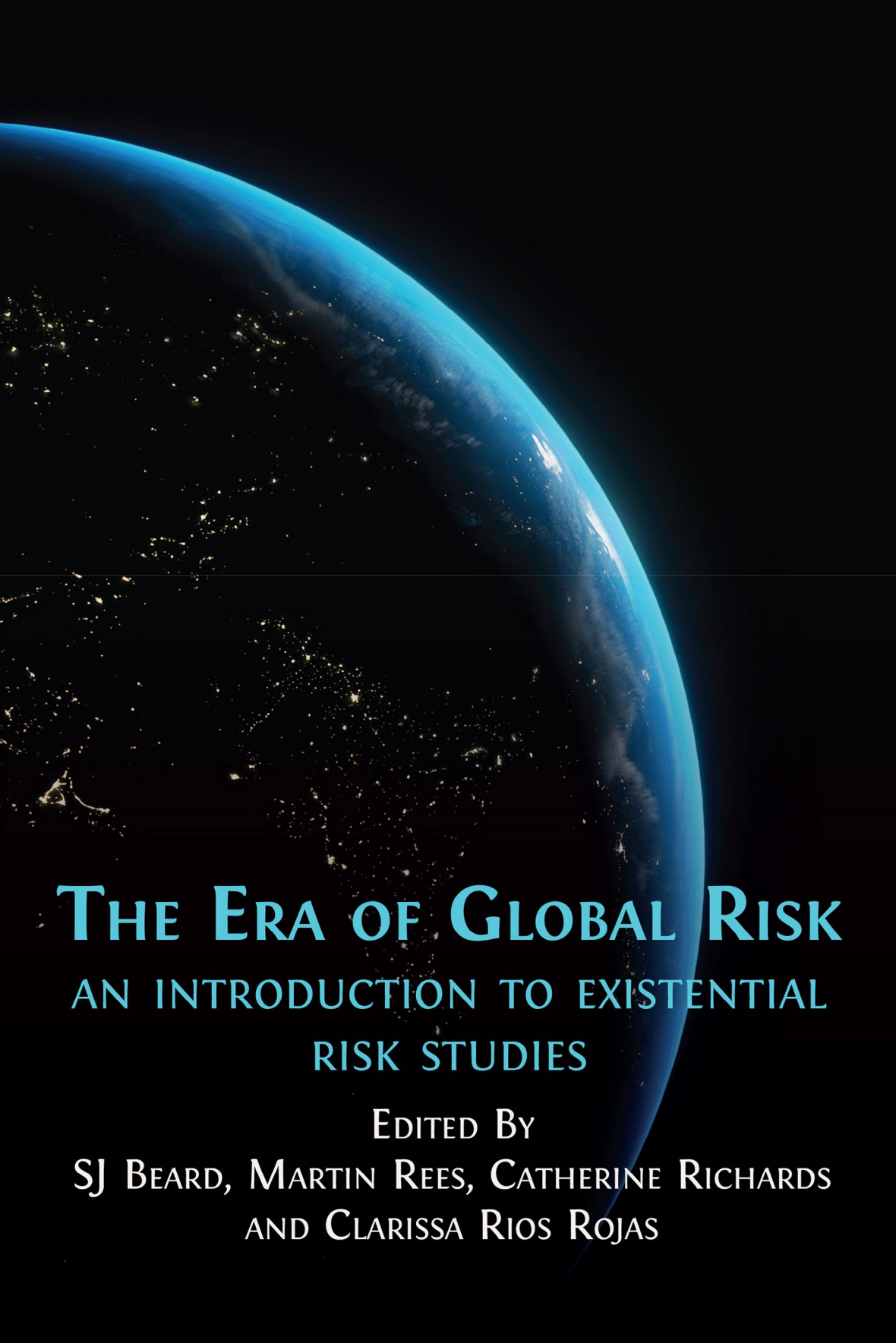

The Era of Global Risk

An Introduction to Existential Risk Studies

Edited By
SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas

SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas (eds), *The Era of Global Risk: An Introduction to Existential Risk Studies*. Cambridge, UK: Open Book Publishers, 2023, https://doi.org/10.11647/OBP.0336

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0336#resources

ISBN Paperback: 978-1-80064-786-2
ISBN Hardback: 978-1-80064-787-9
ISBN Digital (PDF): 978-1-80064-788-6
ISBN Digital ebook (epub): 978-1-80064-789-3
ISBN XML: 978-1-80064-791-6
ISBN HTML: 978-1-80064-792-3
DOI: 10.11647/OBP.0336

Cover image: Anirudh, *Our Planet* (October 14, 2021), https://unsplash.com/photos/Xu4Pz7GI9JY. Cover design by Jeevanjot Kaur Nagpal.

10. Military Artificial Intelligence as a Contributor to Global Catastrophic Risk

Matthijs M. Maas, Kayla Lucero-Matteucci, and Di Cooke

It should hardly be surprising that military technologies have featured prominently in public discussions of global catastrophic risk (GCR).[1] The prospect of uncontrolled global war stands as one of the oldest and most pervasive scenarios of what total societal disaster would look like. Conflict has always been able to devastate individual societies; in the modern era, technological and scientific progress has steadily increased the ability of state militaries, and possibly others, to inflict catastrophic violence.[2]

There are many technologies with this capacity, with artificial intelligence (AI) becoming a more notable one in recent years. Increasingly, experts from numerous fields have begun to focus on AI technologies' applications in warfare, considering how these could pose risks, or even new GCRs. While the technological development of military AI and the corresponding study of its impacts are still at an early stage, both have also progressed dramatically in the past decade. Most visibly, the development and use of Lethal Autonomous Weapons (LAWS) has sparked a heated debate, spanning both academic and political spheres.[3] However, in actuality, military applications of AI technology extend far beyond controversial 'killer robots'—with diverse uses from logistics to cyberwarfare, and from communications to training.[4]

 https://doi.org/10.11647/OBP.0336.10

It is anticipated that these applications may lead to many novel risks for society. The growing trend of utilising AI across defence-related systems creates new potential points for technical failure or operator errors; it can result in unanticipated wide-scale structural transformations in the decision environment or may negatively influence mutual perceptions of strategic stability, exacerbating the potential for escalation resulting in global catastrophic impacts. Even in less directly kinetic or lethal roles, such as intelligence-gathering or logistics, there is concern that the use of AI systems might still circuitously lead to GCRs. Finally, there are possible GCRs associated with the future development of more capable AI systems, such as artificial general intelligence (AGI); while these final potential GCRs are not the direct focus of this chapter, it should be noted that these risks could be especially significant in the military context, and that this would require caution rather than complacency.

Despite the ongoing endeavours around the world to leverage more AI technology within the national security enterprise, current efforts to identify and mitigate risks resulting from military AI are still very much nascent. At a technical level, one of the most pressing issues facing the AI technical community today is that any AI system is prone to a wide array of performance failures, design flaws, unexpected behaviour, or adversarial attacks.[5] Meanwhile, numerous militaries are devoting considerable time and resources towards deploying AI technology in a range of operational settings. Despite this, many still lack clear ethics or safety standards as part of their procurement and internal development procedures for military AI.[6] Nor have most state actors actively developing and deploying such systems agreed to hard boundaries limiting the use of AI in defence, or engaged in establishing confidence-building measures with perceived adversaries.[7]

It is clear that military AI developments could significantly affect the potential for GCRs in this area, making the exploration of this technological progression and its possible impacts vital for the GCR community. Now that AI techniques are beginning to see real-world uptake by militaries, it is more crucial than ever that we develop a detailed understanding about how military AI systems might be considered as GCRs in their own right, or how they might be relevant contributors to military GCRs. In particular, from a GCR perspective, further attention is needed to examine instances when AI intersects with

military technologies as destructive as nuclear weapons, potentially producing catastrophic results. To enable a more cohesive understanding of this increasingly complex risk landscape, we explore the established literature and propose further avenues of research.

Our analysis proceeds as follows: after reviewing past military GCR research and recent pertinent advancements in military AI, this chapter turns the majority of its focus on LAWS and the intersection between AI and the nuclear landscape, both of which have received the most attention thus far in existing scholarship. First examining LAWS, we assess whether they might constitute GCRs, and argue that while these systems are concerning, they do not yet appear likely to be a GCR in the near term, considering current and anticipated production capabilities and associated costs. We then delve into the intersection of military AI and nuclear weapons, which we argue has a significantly higher GCR potential. We examine the GCR potential of nuclear war, briefly discussing the debates over when, where, and why it could lead to a GCR. Furthermore, after providing recent geopolitical context by identifying relevant converging global trends which may also independently raise the risks of nuclear warfare, the chapter turns its focus to the existing research on specific risks arising at the intersection of nuclear weapons and AI. We outline six hypothetical scenarios where the use of AI systems in, around, or against nuclear weapons could increase the likelihood of nuclear escalation and result in global catastrophes. Finally, the chapter concludes with suggestions for future directions of study, and sets the stage for a research agenda that can gain a more comprehensive and multidisciplinary understanding of the potential risks from military AI, both today and in the future.

Risks from (military) AI within the Global Catastrophic Risks field

Before understanding how military AI might be a GCR, it is important to understand how the GCR field has viewed risks from AI more broadly. Within the GCR field, there has been growing exploration of the ways in which AI technology could one day pose a global catastrophic or existential risk.[8] Such debates generally have not focused much on the military domain in the near term, however. Instead, they often focus

on how such risks might emerge from future, advanced AI systems, developed in non-defence (or, at best, broadly 'strategic') contexts or sectors. These discussions have often focused on the development of Artificial General Intelligence (AGI) systems that would display "the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments"[9] with performance equivalent or superior to a human in many or all domains. These are, of course, not the only systems studied: more recent work has begun to explore the prospects for, and implications of, intermediate 'High-Level Machine Intelligence'[10] or 'Transformative AI'[11]—types of AI systems that would be sufficient to drive significant societal impacts—without making strong assumptions about the architecture, or 'generality', of the system(s) in question.

Whichever term is used, across the GCR field (and particularly in the subfields of AI safety and AI alignment) there has been a long-running concern that if technological progress continues to yield more capable AI systems, such systems might eventually pose extreme risks to human welfare if they are not properly controlled or aligned with human values.[12] Unfortunately, pop-culture depictions of AI have fed some misperceptions about the actual nature of the concerns in this community.[13] As this community notes itself, there is still deep uncertainty over whether existing approaches in AI might yield progress towards something like AGI,[14] or when such advanced systems might be achieved.[15] Nonetheless, they point to a range of peculiar failure modes in existing machine learning approaches,[16] which often display unexpected behaviours, achieving the stated target goals in unintended (and at times hazardous) ways.[17] Such incidents suggest that the safe alignment of even today's machine-learning systems with human values will be a very difficult task;[18] that it is unlikely that this task will become easier if or when AI systems become highly capable; and that even minor failures to ensure such alignment could have significant, even globally catastrophic societal impacts.[19]

However, while the continued investigation of such future risks is critical, these are not strictly the focus of this chapter, which rather looks at the intersection of specifically military AI systems with GCRs, today or in the near-term. Indeed, with only a few exceptions,[20] existing GCR research has paid relatively little attention to the ways in which *military* uses of AI could result in catastrophic risk. That is not to say

that the GCR community has not been interested in studying military technologies in general. Indeed, there have been research efforts to learn from historical experiences with the safe development and responsible governance of high-stakes military technologies, to derive insights for critical questions around the development, deployment, or governance of advanced AI. This research includes (for example) analyses of historical scientific dynamics around (strategically relevant) scientific megaprojects,[21] the plausibility of retaining scientific secrecy around hazardous information,[22] or the viability of global arms control agreements for high-stakes military technologies.[23] Other work in this vein has studied the development of, impacts of, and strategic contestation over previous 'strategic general-purpose technologies' with extensive military applications, such as biotechnology, cryptography, aerospace technology, or electricity.[24] However, these previous inquiries work by analogy, and have neglected to thoroughly examine in detail the object-level question of whether or how existing or near-term military AI systems could themselves constitute a GCR.

Thus far, the predominant focus on military AI as GCRs has been on LAWS, and on nuclear weapons. The former should not be surprising, given the strong resonance of 'killer robots' in the popular imagination. The latter should not be surprising, given that the GCR field's examination of military technologies has its roots in original concerns about nuclear weapons. Indeed, in the past 75 years, long before terms such as GCR or existential risk even came to be, the threat of nuclear weapons inspired a wave of work, study, and activism to reckon with the catastrophic threats posed by this technology.[25] Still, at the present moment, the exploration of how military AI might intersect with or augment the dangers posed by destructive technologies such as nuclear weapons is still in its early stages.

Before delving into military AI as a potential GCR, it is also crucial to first define what we consider to be a GCR. Global catastrophic risks (GCRs) are risks which could lead to significant loss of life or value across the globe, and which impact all (or a large portion) of humanity. There is not yet widespread agreement on what this means exactly, what threshold would count as a global catastrophe,[26] or what the distinction is between GCRs and existential risks. For many discussions within the field of GCR, and for many of the risks discussed in other chapters in this volume, such ambiguity may not matter much, if the potential

risks discussed are so obviously catastrophic in their impacts (virtually always killing hundreds of millions, or even resulting in extinction) that they would undeniably be a GCR. Yet in the domain of military AI (as with other weapons technologies), one may confront potential edge-case scenarios—involving the projected deaths of hundreds of thousands, or even millions, but where it is unclear if this (plausibly) would reach higher.

Within our chapter, we therefore need some working threshold for what constitutes a GCR, even if any threshold is (by its nature) contestable. What is a workable threshold to use here for GCRs? One early influential definition by Bostrom and Cirkovic holds that a catastrophe causing 10,000 fatalities (such as a major earthquake or nuclear terrorism) might not qualify as a global catastrophe, whereas one that "caused 10 million fatalities or 10 trillion dollars' worth of economic loss (e.g., an influenza pandemic) would count as a global catastrophe, even if some region of the world escaped unscathed."[27] However, while there is therefore clear definitional uncertainty, in this chapter we will utilise a lower bound for GCRs that lies in the middle of the range indicated by Bostrom and Cirkovic. To be precise, we understand a GCR to be *an event or series of directly connected events which result in at least one million human fatalities within a span of minutes to several years, across at least several regions of the world*.

To understand whether and in what ways military AI could contribute to GCRs of this level, we next sketch the speed and direction by which this technology has been developed and deployed for military purposes, both historically and in recent years.

Advances in military AI: Past and present

The use of computing and automation technologies in military operations itself is hardly new. Indeed, the history of AI's development has been closely linked to militaries, with many early advances in computing technologies, digital networks, and algorithmic tools finding their genesis in military projects and national strategic needs.[28] During the Cold War, there were repeated periods of focus on the military applications of AI, from early RAND forecasts exploring long-range future trends in automation[29] to discussions of the potential use of AI in

nuclear command and control (NC2) systems management.[30] As such, military interest in AI technology has proven broadly robust, despite periods of occasional disillusionment during the 'AI winters'. Even when individual projects failed to meet overambitious goals and were cancelled or scaled back, they still helped advance the state of the art; such was the case with the US's 1980s Strategic Computing Initiative—a ten-year, $1 billion effort to achieve full machine intelligence.[31] Moreover, by the 1990s, some of these investments seemed to be beginning to pay off on the battlefield: for instance, during the first Gulf War, as a wide range of technologies contributed to a steeply one-sided Coalition victory over Iraqi forces,[32] the US military's use of the Dynamic Analysis and Replanning Tool (DART) tool for automated logistics planning and scheduling was allegedly so successful that DARPA claimed this single application had promptly paid back 30 years of investment in AI.[33]

This long-standing relation between militaries and AI technology also illustrates how—just as there is not a single 'AI' technology, but rather a broad family of architectures, techniques, and approaches—likewise there is not one 'military AI' use case (e.g. combat robots). Rather, weapons systems have, for a very long time, been positioned along a spectrum of various forms of automatic, automated, or autonomous operation.[34] Many of these are therefore not new to military use: indeed, armies have been operating 'fire and forget' weapons (i.e. weapons that do not require further external intervention or guidance after launch) for over 70 years, dating back to the acoustic (sound-tracking) homing torpedoes that already saw use during the Second World War.[35] In restricted domains, such as at sea, fully autonomous 'Close-in Weapon Systems' (last-defence anti-missile cannons) have been used for years by dozens of countries to defend their naval vessels.[36]

Still, recent years have seen a notable acceleration in the militarisation of AI technology.[37] The market for the use of AI in military uses was estimated at $6.3 billion in 2020, and was then projected to double to $11.6 billion by 2025.[38] Investments are led by the US, China, Russia, South Korea, the UK, and France,[39] but also include efforts by India, Israel, and Japan.[40]

What is the exact appeal of AI capabilities for militaries? Generally speaking, AI has been described as a 'general-purpose technology' (GPT),[41] suggesting that it is likely to see global diffusion and uptake, even if there may be shortfalls amid rushed applications.[42] This also extends

to the military realm. Although uptake of military AI differs by country, commonly highlighted areas of application include improved analysis and visualisation of large amounts of data for planning and logistics; pinpointing relevant data to aid intelligence analysis; cyber defence and identification of cyber vulnerabilities (or, more concerningly, cyber offence); early warning and missile defence; and autonomous vehicles for air, land, or sea domains.[43]

Given this range of uses, there has been significant government attention for the strategic promise of the technology. US scholars describe AI as having prompted a new 'revolution in military affairs';[44] Chinese commentators project that virtually any aspect of military operations might be improved, made faster, or more accurate—or as they call it, 'intelligentised'[45]—through AI. In this way, AI could enable 'general-purpose military transformations' (GMT).[46] Consequently, many anticipate far-reaching or even foundational changes in military practice. Even those with a more cautious outlook still agree that AI systems can serve as a potent 'evolving' and 'enabling' technology that will have diverse impacts across a range of military fields.[47] This has led some to anticipate widespread and unconstrained proliferation of AI, on the assumption that "[t]he applications of AI to warfare and espionage are likely to be as irresistible as aircraft".[48] Still, this should come with some caveats.

In the first place, many applications of military AI may appear relatively 'mundane' in the near term. As argued by Michael Horowitz, "[m]ost applications of AI to militaries are still in their infancy, and most applications of algorithms for militaries will be in areas such as logistics and training rather than close to or on the battlefield".[49] Indeed, early US military accounts on autonomy maintain that there are only particular battlefield conditions under which that capability adds tactical value.[50] Despite the ambitious outlook and rhetoric of many national defence strategies around AI, in practice their focus appears to be more on rapidly maximising the benefits from easily accessible or low-hanging AI applications in areas such as logistics and predictive maintenance, rather than working immediately towards epochal changes.[51]

Secondly, while there are significant technological breakthroughs in AI, a number of technological and logistical challenges are likely to slow implementation to many militaries, at least in the near

term of the next decade. All military technologies, no matter how powerful, face operational, organisational, and cultural barriers to adoption and deployment,[52] and there is no reason to expect military AI will be immune to this. Indeed, militaries may face additional and unexpected hurdles when forced to procure such systems from private-sector tech companies, because of mismatches in organisational processes, development approaches, and system requirements,[53] or export control restrictions or military robustness expectations that go beyond consumer defaults.[54] Finally, emerging technologies, when in their early stages of development, will often face acute trade-offs or brittleness in performance that limit their direct military utility.[55] The often high-profile failures of—or accidents with—early systems can also temper early military enthusiasm for deployment, stopping or slowing development, especially where it concerns more advanced applications such as complex drone swarms with the capacity for algorithmically coordinated behaviour.[56]

Moreover, there are factors that may slow or restrict the proliferation of military AI technology, at least in the near term. Military technological espionage or reverse engineering has proven a valuable but ultimately limited tool for militaries to keep pace with cutting-edge technologies developed by adversaries.[57] In recent years, the training of cutting-edge AI systems has also begun to involve increasingly large computing hardware requirements,[58] as well as important AI expert knowledge, which could ultimately restrict the straightforward proliferation of many types of military AI systems around the globe.[59]

Finally, and alongside all of this, there may be political brakes, or even barriers, to some (if not all) military uses of AI. It should be kept in mind that while the adoption of any military technology may be driven by military-economic selection pressures,[60] their development or use by any actors is certainly not as inevitable or foregone as it may appear in advance.[61] Historically, states and activists have—by leveraging international norms, interests, and institutions—managed to slow, contain, or limit the development of diverse sets of emerging weapons technologies (from blinding lasers to radiological weapons, and from environmental modification to certain nuclear programs), achieving successes that, while not always perfect, often exceeded initial expectations.[62] Accordingly, there is always the possibility that

the coming decades will see invigorated opposition to military AI that will impose an effective brake; however, the success of any such efforts will depend sensitively on questions of issue-framing, forum choice, and organisation.[63]

As a result, the reality of military AI may appear relatively mundane, at least for the next few years, even as it gathers pace below the surface. Nonetheless, even under excessively conservative technological assumptions—where we assume that AI performance progress slows down or plateaus in the next years—AI appears likely to have significant military impacts. In fact, in many domains, it need not achieve further dramatic breakthroughs for existing capabilities to alter the international military landscape. As with conventional drone technologies, even imperfect AI capabilities (used in areas such as image recognition) could suffice to enable disruptive tactical and strategic effects, especially if they are pursued by smaller militaries or non-state actors.[64] As such, even if we assume that more advanced AI capabilities remain out of reach or undesired (an assumption that may rest on thin ground), the development of autonomous systems could herald a wide range of tactical changes,[65] including a shift in the so-called 'offense-defence balance'[66] due to increased effectiveness of offensive capabilities—along with an increased use of deception and decoys, or changes in force operation and operator skill requirements, to name a few.[67] But the question still remains: are any of these impacts plausibly globally catastrophic?

LAWS as GCRs

Thus far, some of the most in-depth discussions of military AI systems as plausible GCRs have focused on the potential risks of LAWS. In this section, we examine existing research and explore several proposed scenarios for ways by which LAWS might contribute to GCRs. Ultimately, we argue that the threshold of destruction (>one million human fatalities) necessary for a GCR leaves most (if not all) near-term LAWs unlikely to qualify as GCRs in isolation.

To pose a GCR, a technology must, at some point, have lethal effects. To be certain, there are significant developments in directly lethal military AI. Of course, technical feasibility by itself does not mean the

development of such systems is inevitable: the existence of LAWS—or their mass procurement and deployment beyond prototypes—hinges not just on questions of technological feasibility, but also on questions of governments' willingness to deploy such systems. To take the technological developments as a starting point, LAWS systems are already being developed and deployed across militaries worldwide. Already in 2017, a survey identified "49 deployed weapon systems with autonomous targeting capabilities sufficient to engage targets without the involvement of a human operator".[68] This number has grown substantially since.

Moreover, in the past years the first fully autonomous weapons systems have reportedly begun to see actual (if limited) deployment. For instance, the South Korean military briefly deployed Samsung SGR-A1 sentry gun turrets to the Korean Demilitarised Zone, which came with an optional autonomous operation mode.[69] Israel has begun to deploy the 'Harpy' loitering anti-radar drone,[70] and various actors have begun to develop, sell, or use weaponised drones capable of autonomy.[71] In 2019, the Chinese company Ziyan released the Blowfish A3: a machine-gun-carrying assault drone that was allegedly marketed as sporting 'full autonomy'.[72] 2020 saw claims that Turkey had developed (semi-)autonomous versions of its 'Kargu-2' kamikaze drone;[73] in the spring of 2021, a UN report suggested that this weapon had been used fully autonomously in the Libyan conflict, to attack soldiers fleeing battle.[74] UAVs that are, in principle, capable of full autonomy have also reportedly seen use in the 2022 Russian invasion of Ukraine, although it remains difficult to ascertain whether any of these systems have been used in fully autonomous mode.[75] Recent developments in autonomous weapons have also included the use of large numbers of small robotic drone platforms in interacting swarms.[76] The Israel Defense Forces deployed such swarms in the May 2021 campaign on Gaza: to locate, identify, and even strike targets.[77]

In other cases, AI has been used in ways that are less autonomous, but which certainly show the lethality-enabling function of many AI technologies.[78] For example, the November 2020 assassination of Mohsen Fakrizadeh (Iran's top nuclear scientist) relied upon a remotely controlled machine gun. While the system was controlled by a human operator, it reportedly used AI to correct for more than a

second-and-a-half of input delay. This allowed the operator to fire highly accurately at a moving target, from a moving gun platform on a highway, while stationed more than 1,000 miles away.[79] Other developments demonstrate the potential for more advanced autonomous behaviour. In 2020, DARPA ran AlphaDogFight, a simulated dogfight between a human F-16 pilot and a reinforcement-learning-based AI system, which saw the AI defeating the human pilot in all of their five matches.[80] In the past decade, the US and others have also experimented with a plane-launched swarm of 103 Perdix drones, which coordinated with one another to demonstrate collective decision-making, adaptive formation, and 'self-healing' behaviour.[81] Experiments in swarming drones have continued apace since.

Perhaps unsurprisingly—due to the fact that it has had earlier adoption relative to other high-risk military applications—LAWS have received sustained public scrutiny and scholarly attention, far more so than any other military AI use case. Consequently, efforts to develop governance approaches have arisen from multiple corners,[82] including at the UN Convention on Certain Conventional Weapons (CCW) since 2014, as well as within arms control communities since 2013.[83] However, it is notable that these debates have mostly examined qualitative characteristics of LAWS, rather than the potential quantitative upper limit on the scale of violence they might enable. Specifically, opposition to LAWS has focused primarily (but not exclusively) on their potential violation of various existing legal principles or regimes under international law, specifically International Humanitarian Law,[84] or (when used in law enforcement outside of war zones) under international human rights law;[85] other discussions have explored whether LAWS, even if they narrowly comply with cornerstone IHL principles, might still be held to undermine human dignity because they involve 'machine killing'.[86]

Over time, however, some civil society actors have begun to attempt to understand and stigmatise LAWS swarms as a potential 'weapon of mass destruction',[87] with swarms of lethal drones as a weapon system that could easily fall in the hands of terrorist actors or unscrupulous states, allowing the infliction of massive violence. This is a framing that has become more prominent within counter-LAWS disarmament campaigns,[88] most viscerally in depictions of terror attacks using fully

autonomous microdrones that deliver small, shaped charges (such as the Future of Life Institute's 'Slaughterbot' campaigns of 2017 and 2021).[89] This is indicative of a growing concern for the 'quantitative' dimension and potential scale of mass attacks using autonomous weapons.

As a consequence, two distinct scenarios have often been proposed regarding LAWS technology as a significant global risk: terrorist use for mass attacks and state military use of massed LAWS forces.

Mass terror attacks on public or on GCR-sensitive targets

One hypothetical discussed by experts focuses on the use of LAWS not by state militaries, but by non-state actors (such as terror groups).[90] In theory, terrorists could subsequently leverage larger and larger swarms, either through direct acquisition of such militarised technology (if unregulated), or remote subversion of existing fleets using cyberattacks. Turchin and Denkenberger argue that increasingly larger quantities of drone swarms would be feasible as a global catastrophic risk, as it becomes cheaper to build drones.[91] While it is possible that this could enable mass-casualty attacks, it seems unlikely that any non-state actor could scale such attacks up to the global level. Moreover, it would be hard for them to prepare attacks of such magnitude undetected.

Another less explored risk would involve the (terrorist) use of LAWS to deliver other GCR-capable weapons or agents. For instance, Kallenborn and Bleek have suggested that actors could use drone swarms to deliver existing chemical, biological, or radiological weapons;[92] others have suggested that non-state actors could refit crop-duster drones to disperse chemical or biological agents.[93] In such cases, the level of risk is less clear: it might still be unlikely that these hypothetical events could be scaled up to result in a full GCR; however, this depends on the potency of the delivered agent in question. Ultimately, existing research is still very preliminary, and much further research is necessary to enable more concrete conclusions.

A third attack pathway could involve the malicious or terrorist use of autonomous weapons on sensitive critical infrastructures which, if damaged or compromised, would precipitate GCRs (or at least would instantly cripple our ability to respond to ongoing or imminent GCRs). Drone systems have been used by various non-state actors in recent

years to mount effective attacks against critical infrastructures—as in the attacks on oil pipelines and national airports in the Yemen conflict.[94] Moreover, across the world there are a wide range of vulnerable global infrastructural 'pinch points' (internet connection points, narrow shipping canals, breadbasket regions) which, if they are attacked or degraded, could precipitate major shocks in the global system.[95] Many of these could be conceivably attacked through autonomous weapons, which could result in regional or even global disaster by the resulting knock-on effects, even if they were only temporarily disrupted. For instance, AWS could be used to deliver coordinated attacks on nuclear power plants, potentially resulting in large fallout patterns and contamination of land and food.[96] Alternatively, they could be used to attack and interrupt any future geo-engineering programs, potentially triggering climatic 'termination shocks' (where temperatures bounce back in ways that would be catastrophically disruptive to the global ecosystem and agriculture).[97] However, these types of attack do not seem to necessarily require autonomous weapons, and while they could certainly result in widespread global chaos, it is again unclear if they could be scaled up to the threshold of a global catastrophe involving over one million casualties.

State attacks with massed LAWS swarms

Within existing research, another frequently discussed hypothetical scenario is the idea of well-resourced actors using mass swarms of LAWS to carry out global attacks, allowing for "armed conflict to be fought at a scale greater than ever".[98] There is also a lively discussion about the possibility that mass attacks using swarms of 'slaughterbots' could allow small-state actors to mount attacks that would kill as many as 100,000 people.[99]

Turchin and Denkenberger have argued that in large enough quantities, drone swarms could be destructive enough to constitute a GCR, and command errors could result in autonomous armies creating a similar level of damage. Still, they predict that, even in those scenarios, LAWS are likely to result in broad instability rather than destruction on the scale of a GCR.[100] More recently, Anthony Aguirre has suggested that mass swarms of 'anti-personnel AWS' could deliver large-scale

destruction at lower costs and lower access thresholds than would be required for an equivalently destructive nuclear strike (of an equivalent scale as the Hiroshima bombing), and that such weapons could be scaled up to inflict extreme levels of global destruction.[101] Turchin has suggested that drone swarms could become catastrophic risks only under very specific conditions, where more advanced (e.g. AGI) technologies are delayed, drone manufacture costs fall to extremely low bounds, defensive counter-drone capabilities lag behind, and militaries adopt global postures that condone the development of drone swarms as a strategic offensive weapon.[102] Even under these conditions, he suggests, drone swarms would be unlikely to ever rise to the level of an existential risk, though they could certainly contribute to civilisational collapse in the event of an extensive global war.[103]

Evaluating the feasibility of mass LAWS swarm-attack scenarios as GCRs

In both of the above cases, there is reason for concern and precautionary study and policy. However, there remain at least some practical reasons to doubt that LAWS lend themselves to precipitating catastrophes at a full GCR scale in the near term.

For one, it still is unclear if LAWS would be more cost-effective as a mass-attack weapon for states that have other established options. On the one hand, Aguirre has argued that 'slaughterbots' could be as inexpensive as $100, meaning that, even with a 50% unit attack success rate, and a doubling of cost to account for delivery systems, the shelf price of an attack inflicting 100,000 casualties would be $40 million.[104] However, how does that actually compare to the costs of other mass-casualty weapons systems? While precise procurement costs remain classified, estimates have been given for various nuclear weapon assets: US B61 gravity bombs are estimated to cost $4.9 million each (with a B-52H bomber carrying 20 such bombs costing an additional $42 million); a Minuteman III missile costs $33.5 million apiece (or $48.5 million, including the cost of three nuclear warheads).[105] The cost of North Korean nuclear weapons has been estimated at between $18 million and $53 million per warhead.[106] Accurate and up-to-date cost-effectiveness estimates for other weapons of mass destruction

are hard to come by—in 1969, a UN study estimated that the costs of inflicting one civilian casualty per square kilometre were about $2,000 with conventional weapons, $800 with nuclear weapons, $600 with chemical weapons, and only $1 with biological weapons.[107] However, these estimates are likely considerably outdated, and are unlikely to reflect the destructive efficiency of contemporary WMDs used against modern societies. So, in principle (and perceived only from a narrowly economic perspective), LAWS swarms might appear less cost-effective than most existing WMDs, although not dramatically so. Even then, such swarms could theoretically be competitive, because they are seen as more accessible or achievable than other WMDs (in the sense that their production may be less reliant on globally controlled resources such as fissile materials or toxins).

Moreover, there may be supply-chain limitations, which could result in caps on how many such drone swarm units could be plausibly produced or procured. To be sure, assuming very small drones, swarms could be scaled up to hundreds of thousands or millions of units. Some accounts of drone swarms have envisaged a future of 'smart clouds' of billions of tiny, insect-like drones.[108] Yet this might trade off against effective lethality: it seems unlikely that micro-drone systems will be able to do much more than reconnaissance, given limits in terms of power, range, processing, and/or payload capacity.[109] By contrast, focusing on LAWS that are able to project lethal force at meaningful ranges, the production constraints seem more serious. We can compare the production lines for military drones, a technology with more well-established supply chains: a 2019 estimate by defence information group Janes estimated that more than 80,000 surveillance drones and 2,000 attack drones would be purchased around the world in the next decade.[110] The civilian drone market is admittedly larger, with around five million consumer drones being sold in 2020—a number expected to rise to 9.6 million by 2030.[111]

This suggests that if commercial supply chains were all dedicated to the production of LAWS, GCR-scale attacks could come into range. Yet the relatively small size of the military drone market is still suggestive of the challenges around procuring sufficient numbers of autonomous weapons to truly inflict global catastrophe in the next decade or so, and possibly beyond.

Of course, there might also be counter-arguments that suggest these barriers could be overcome, making mass LAWS attacks (at GCR scale) more feasible. For instance, it could be misleading to look at the raw number of platforms acquired and deployed, since individual autonomous weapons platforms might easily be equipped with weapons that would allow each platform to kill not one but dozens or thousands, depending on the weapon delivered or location of attacks. However, this is not the way that 'slaughterbots' are usually represented; indeed, outfitting these systems with more ordnance would simply make the ordnance the bottleneck.

In the second place, motivated states might be able to step up production and procure far larger numbers of these systems than is possible today, especially if the anticipated strategic context of their use is not counterinsurgency but a near-peer confrontation, where drone swarms might become perceived (either accurately or not) as not just helpful or cost-saving, but also providing a key margin of dominance. For instance, the US Navy in 2020 discussed offensive and defensive tactics for dealing with attacks of 'super swarms' of up to a million drones.[112] Increased state attention and enthusiasm for this technology could change the industrial and technical parameters rapidly.

In the third place, economies of scale and advances in manufacturing capabilities could mean that unit production costs could fall, or mass production could be facilitated, potentially enabling the targeting of many millions. It is unclear to what level costs would have to fall for GCR-scale fleets to become viable (let alone common), however, with Turchin suggesting unit costs of below $1.[113] Even so, barring truly radical manufacturing breakthroughs, producing this would require quite significant investments. The above does not even begin to address questions of delivery.

The overall point here is therefore not that states will remain disinterested in—or incapable of building—drone swarms of a size that would enable GCR-scale attacks. Indeed, states have often proven willing to invest huge sums in military technologies and their production infrastructures and industries.[114] Still, even in those cases, LAWS swarms will likely not be as destructive as modern thermonuclear weapons: as argued by Kallenborn, "[w]hile they are unlikely to achieve the scale of harm as the Tsar Bomba, the famous Soviet hydrogen bomb, or most

other major nuclear weapons, swarms could cause the same level of destruction, death, and injury as the nuclear weapons used in Nagasaki and Hiroshima."[115] That suggests that they might be seen by militaries to complement rather than substitute for existing deterrents.

The above suggests that LAWS are certainly a real concern, in that it appears possible that, if this technology is developed further, it could in principle be used to inflict mass-casualty attacks on cities; at the same time, it implies that unless political, economic, or technological conditions change, swarms of LAWS (whether operated by terrorists or states) remain unlikely to be able to inflict GCR-level catastrophes in the near future. The scale-up that would be necessary to achieve destruction that would qualify it as a GCR does not presently seem to be a realistic outcome, both industrially but also politically—particularly given the host of similarly or more destructive weapons already available to states. All this suggests that, while autonomous weapons would likely be disruptive, their use would not scale up to a full GCR under most circumstances. Nevertheless, there may be additional edge cases of risk, especially in the under-explored scenarios such as the use of LAWS to deliver WMDs, and/or their use in mass-scale internal repression or genocide.[116] This, therefore, is an area that will require further research.

Nuclear weapons and AI

There is a second way in which military AI systems could rise to become a GCR: this is through their interaction with one of the oldest anthropogenic sources of global catastrophic risk: nuclear weapons.

Nuclear war as a GCR

To understand the way that AI systems might increase the risk of nuclear war in ways that could pose GCRs, it is first key to briefly review the ways in which nuclear war itself has become understood as a global catastrophic risk.

Since the invention of atomic weapons, discussions of nuclear risk have often been characterised by sharply divergent frames and understandings, with many accounts focusing single-mindedly either on the perceived irreplaceable strategic and geopolitical benefits derived from possessing nuclear weapons, or on the absolutely intolerable

humanitarian consequences of their use. The discourses surrounding nuclear weapons today often still fall within those categories.[117] This is not new: early understandings of nuclear weapons vacillated between treating them as simply another weapon for tactical use on the battlefield,[118] or as an atrocious weapon of genocide,[119] potentially even capable of incinerating the atmosphere, as some lead Manhattan Project scientists briefly worried might happen during the Trinity test.[120]

One fact which no one questions, however, is the historically unprecedented capability of nuclear weapons to inflict violence at a massive scale.[121] The crude atomic bombs dropped by the US on Hiroshima and Nagasaki killed at least 140,000 and 74,000 people respectively, but more recently, nuclear weapons with similar destructive capacity have been considered 'low-yield'.[122] In the decades following the Second World War, countries developed thermonuclear weapons which, in some cases, were thousands of times more destructive than the first atomic bombs.[123] Today, the use of a single nuclear weapon could kill hundreds of thousands of people, and a nuclear exchange—even involving 'only' a few dozen nuclear weapons—could have devastating consequences for human civilisation and the ecosystems upon which we depend.[124]

If the use of a single nuclear weapon would be a tragedy, the additional fact that these weapons would rarely be used in isolation highlights clear paths to global catastrophe. According to David Rosenberg, early US plans for a nuclear war (drawn up by the Strategic Air Command in 1955) were estimated to be able to inflict a total of 60 million deaths and another 17 million casualties on the Soviet Union.[125] Later plans would escalate even further. The 1962 US nuclear war plan, utilising the entire US arsenal, would have killed an estimated 285 million people and harmed at least another 40 million in the targeted (Soviet-Sino bloc) countries alone.[126] Daniel Ellsberg, then at DARPA, later recounted war plans for a US first-strike on the Soviet Union, Warsaw Pact satellites, and China, as well as additional casualties from fallout in adjacent neutral (or even allied) countries, which projected global casualties rising up to 600 million.[127]

These estimates proved not to be a ceiling but a potential lower bound, once scientists began to focus on potential environmental interactions of nuclear war. In 1983, Carl Sagan famously embarked on a public campaign to raise awareness about the environmental

impacts of nuclear weapons. Along with several colleagues, including some in the USSR, Sagan disseminated a theory of "nuclear winter", which holds that fires caused by nuclear detonations would loft soot into the stratosphere, leading to cooler conditions, drought, famine, and wide-scale death.[128] In response to Sagan's campaign, the US government attempted to downplay public discussions of nuclear winter, with the Reagan administration stating publicly in 1985 that it had "...very little confidence in the near-term ability to predict this phenomenon quantitatively."[129] Still, archival materials reveal that, internally, administration officials had strong feelings about nuclear winter. One employee of the Department of Defense noted at the time that the US government and overall scientific community "ought to be a bit chagrined at not realizing that smoke could produce these effects."[130]

Over time, accounts such as these have led to the creation of a nuclear taboo, or norm of non-use,[131] although it is unclear whether the taboo will stand amid a number of contemporary developments.[132] Today, scholars continue to study the impacts of nuclear detonations, with some predicting that even a small nuclear exchange could result in nuclear winter. For instance, climate scientist Alan Robock and colleagues suggest that "...if 100 nuclear bombs were dropped on cities and industrial areas—only 0.4 percent of the world's more than 25,000 warheads—[this] would produce enough smoke to cripple global agriculture."[133] Even in the limited scenario of such a 'nuclear autumn', it has been estimated that US and Chinese agricultural production in corn and wheat would drop by about 20–40% in the first five years, putting as many as two billion people at risk of starvation.[134] A larger exchange between the US and Russia would have even more serious and catastrophic consequences, according to a 2019 analysis of long-term climatic effects.[135]

To be sure, there remains some dissent over models predicting these environmental impacts,[136] the science of nuclear winter,[137] or the status of nuclear war as GCR.[138] Assessments of nuclear risk are made more difficult still by uncertainty in not just the environmental models, but also the underlying strategic dynamics. There are deep methodological difficulties around quantifying nuclear risks, especially since an all-out nuclear war has never occurred. Whereas studies of some (but certainly not all) other GCRs, such as pandemics,

can aim and extrapolate from historical disasters, scholars examining the risk of nuclear war face the steep challenge of attempting to "understand an event that never happened".[139] Nonetheless, different approaches attempt to integrate historical base rates for intermediary steps (close calls and accidents) with expert elicitation, to come to imperfect background estimates.[140]

Yet (as even modellers note), such estimates remain subject to extreme uncertainty, given the unpredictability of strategy, targeting decisions, and complex socio-technical systems. A host of close calls during the Cold War show that carefully designed systems are not impervious to accidents or immune from human error.[141] As normal accident theory suggests, undesirable events and accident cascades are inevitable,[142] and adding in automated components or fail-safe systems may sometimes counterintuitively increase overall risk by increasing the system's complexity, reducing its transparency, or inducing automation bias.[143] The present era is now faced with the question of whether emerging technologies such as AI will be equally susceptible to risks from normal accidents,[144] whether they will contribute to such risks in legacy technologies such as nuclear weapons, and whether they will make the impacts of already destructive weapons more severe or increase the likelihood of their use.

Overall, the massive loss of life envisioned in nuclear war plans certainly qualifies nuclear weapons as a GCR. Whether they are considered to pose an existential risk may depend on the number and yield of weapons used. Some analyses have suggested that, even in extreme scenarios of nuclear war that resulted in civilisational collapse and the deaths of very large (>90% or >99.99%) fractions of the world population, we might still expect humanity to survive.[145] On the other hand, it has been countered that, even if such a disaster would not immediately lead to extinction, it might still set the stage for a more gradual and eventual collapse or extinction over time, or at the very least for the recovery of a society with much worse prospects.[146] However, for many commonly shared ethical intuitions, this distinction may be relatively moot.[147] Whether or not it is a technical existential risk, any further study of nuclear weapons' environmental and humanitarian impacts, including nuclear winter, will likely further corroborate their status as a major threat to humanity both today and into the future.

Recent developments in nuclear risk and emerging technology

Today's emergence of military AI therefore comes on top of a number of other disruptive developments that have already impacted nuclear risk over the past decades, and which have already brought concern about nuclear GCRs to the forefront.

Notably, this attention comes after a period of relative inattention to nuclear risk. In the aftermath of the Cold War, the risks posed by the existence of nuclear weapons were seen to be less immediate and pronounced. Accordingly, discussions came to focus more on nuclear security, including efforts after the fall of the Berlin Wall to secure Soviet nuclear materials,[148] as well as the challenges of preventing terrorist acquisition of WMDs, such as through the UNSC Resolution 1540 and the Nuclear Security Summit initiatives. In the last decade, however, converging developments in geopolitics and military technology have brought military (and especially nuclear) GCRs back to the fore.

First, the relative peace that followed the Cold War has been replaced by competition between powerful states, rather than fully cooperative security (or hegemony) in many domains. Geopolitical tensions between major powers have been inflamed, visible in the form of flashpoints from Ukraine to the South China Sea. Meanwhile, the regimes for the control of WMDs have come under pressure.[149] Nuclear arms control agreements between the US and Russia (such as the Intermediate-Range Nuclear Forces Treaty and the Anti-Ballistic Missile Treaty) have been cancelled by Presidents Trump and Bush; other nuclear states such as the UK, France, or China are not restrained by binding nuclear arms control agreements. Although the US and Russia extended the New Strategic Arms Reduction Treaty in March of 2021,[150] the future of arms control is uncertain amid ongoing disputes between the owners of the world's two largest nuclear arsenals,[151] and tensions between the West and Russia over Putin's invasion of Ukraine. In the absence of open channels of communication and risk reduction measures, the dangers of miscalculation are pronounced.[152]

Second, various states have undertaken programs of nuclear re-armament that reach beyond maintenance and replacement of existing systems, opposing the spirit of the Nuclear Non-Proliferation Treaty's commitment to continued disarmament.[153] For example, the

US recently deployed a new low-yield submarine-launched ballistic missile and requested funding for research and development on a new sea-launched cruise missile.[154] Seeing its nuclear arsenal as guarantor of its great-power status, Russia has modernised its nuclear arsenal,[155] as well as investing in a new generation of exotic nuclear delivery systems, including Poseidon (autonomous submarine nuclear drones),[156] Burevestnik (nuclear-powered cruise missile),[157] Kinzhal (air-launched ballistic missile), and Avangard (hypersonic glide vehicle).[158] While the Chinese nuclear force still lags substantially behind those of its rivals in size, it too has begun a program of nuclear force expansion; analysts estimate that its arsenal has recently surpassed France's to become the world's third largest,[159] and there are concerns that the construction of new ICBM fields shows an expansion in force posture from minimum to medium deterrence.[160] China in 2021 also conducted an alleged test of a Fractional Orbital Bombardment System (FOBS).[161] In its 2021 Integrated Review, the UK recommended an expansion of its nuclear stockpile by over 40%, to 260 warheads.[162]

The third trend relates to the ways in which strategic stability is further strained by the introduction of new technologies, from the United States' Conventional Prompt Global Strike to a range of programs aimed at delivering hypervelocity missiles, which risk exacerbating nuclear dangers by shortening decision timelines, or which introduce 'warhead ambiguity' around conventional strikes which could be mistaken as nuclear ones.[163] New technologies will make states more adept at targeting one another's nuclear arsenals, creating a sense of instability that could lead to pre-emption and/or arms-racing.[164] Not only are states engaging individually in the development of these technologies, the last few years have also seen an increasing number of strategic military partnerships involving such technologies, and shaping and constraining their use.[165]

In sum, there are several external trends that frame the historical intersection of nuclear risk with emerging military AI technologies: an increase in inter-state geopolitical tensions, state nuclear rearmament or armament, and the introduction of other novel adjacent technologies. These trends all intersect with the advances of military AI, and against the backdrop of an alleged 'AI Cold War'.[166]

This brings us back to our preceding discussion: even if many military AI applications are not a direct GCR, there are concerns at their intersection with nuclear weapons. Yet how, specifically, could the use of AI systems to automate, support, attack, disrupt, or change nuclear decision-making interact with the already complex geometry of deterrence, creating new avenues for deliberate or inadvertent global nuclear catastrophe?

Nuclear weapons and AI: Usage and escalation scenarios

As discussed, militaries have a long history of integrating computing technologies with their operations—and strategic and nuclear forces are no exception. This has led some to raise concerns about the potential risks of such integrations. In the late 1980s, Alan Borning noted that "[g]iven the devastating consequences of nuclear war, it is appropriate to look at current and planned uses of computers in nuclear weapons command and control systems, and to examine whether these systems can fulfil their intended roles".[167] On the Soviet side, there were similar concerns over the possibility of triggering a 'computer war', especially in combination with launch on warning postures and the militarisation of space. As Soviet scholar Borish Raushenbakh noted, "[t]otal computerization of any battle system is fraught with grave danger".[168] Scruples notwithstanding, during the late Cold War the Soviet Union did in fact develop and deploy the 'Perimeter' (or 'Dead Hand') system; while still including a small number of human operators, when switched on during a crisis period the system was configured to (semi-)automatically launch the USSR's nuclear arsenal, if its sensors detected signs of a nuclear attack and lost touch with the Kremlin.[169]

As previously stated, concerns about the potentially escalatory effects of AI on the nuclear landscape have been somewhat more extensively examined than other possible military AI GCR scenarios. In this section, we examine established research investigating potential risk scenarios arising from the intersection between AI and the nuclear weapons infrastructure. We therefore concern ourselves not only with the direct integration of AI into nuclear decision-making functions, such as launch orders, but also with the application of AI in supporting

or tangentially associated systems, as well as its indirect effects on the broader geopolitical landscape. Throughout the Cold War, US and Soviet NC3 featured automated components, but today there is an increasing risk that AI will begin to erode human safeguards against nuclear war. Although NC3 differs by country, we define it broadly as *the combination of warning, communication, and weapon systems—as well as human analysts, decision-makers, and operators—involved in ordering and executing nuclear strikes, as well as preventing unauthorised use of nuclear weapons.*

NC3 systems can include satellites, early warning radars, command centres, communication links, launch control centres, and operators of nuclear delivery platforms. Depending on the country, individuals involved in nuclear decision-making might include operators of warning radars, analysts sifting through intelligence to provide information about current and future threats, authorities who authorise the decision to use nuclear weapons, or operators who execute orders.[170] Differences in posture among nuclear weapon possessors mean that their NC3 varies considerably: for example, while China has dual-use land- and sea-based nuclear weapons,[171] the United Kingdom has only a sea-based nuclear deterrent, and its NC3 systems do not support any conventional operations.[172]

To understand how AI could affect the risk of a global nuclear war, it is important to distinguish between distinct escalation routes. Following a typology by Johnson,[173] we can distinguish intentional and unintentional escalation. Under (1) *intentional* escalation, one state has (or gains) a set of (AI + nuclear) strategic capabilities, as a result of which they knowingly take an escalatory action for strategic gain (e.g. they perceive they have a first-strike advantage, and launch a decapitation strike); this stands in contrast to various forms of (2) *unintentional* escalation—situations where "an actor crosses a threshold that it considers benign, but the other side considers significant".[174]

Specifically, unintentional escalation can be further subdivided into (2a) *inadvertent* escalation (mistaken usage on the basis of incorrect information); (2b) *catalytic* escalation (nuclear war between actors A and B, triggered by the malicious actions of a third party C against either party's NC3 systems); or (2c) *accidental* escalation (nuclear escalation without a deliberate and properly informed launch decision, triggered

by a combination of human and machine interaction failures, as well as background organisational factors).[175]

Additionally, AI can be used in, around, and against NC3 in a number of ways, all of which can contribute to different combinations of escalation risk (and thereby GCR). We will therefore review some uses of military AI, and how these could increase the risk of one or more escalation routes being triggered.

Autonomised decision-making

The first risk involves integrating AI directly into NC3 nuclear decision-making.[176] This could involve giving systems the ability to authorise launches, and/or to allow AI systems to compose lists of targets or attack patterns following a launch order, in ways that might not be subject to human supervision.

It should be immediately noted that few states currently appear interested in the outright automation of nuclear command and control in any serious way.[177] While commentators within the US defence establishment have called for the US to create its own AI-supported nuclear 'Dead Hand',[178] senior defence officials have explicitly claimed they draw the line at such automation, ensuring there will always be a human in the loop of nuclear decision-making.[179] Likewise, Chinese programs on military AI currently do not appear focused on automated nuclear launch.[180]

Indeed, in addition to a lack of interest, there may be outstanding technical limits and constraints posed by existing AI progress. For instance, it has been argued that current machine-learning systems do not lend themselves well to integration in nuclear targeting, given the difficulty of collating sufficient (and sufficiently reliable) training datasets of imagery of nuclear targets (e.g. mobile launch vehicles), which some have argued will provide 'enduring obstacles' to implementation.[181] If that is the case, highly anticipated applications may remain beyond current AI capabilities.

Nonetheless, even if no state is known to have directly done so today, and some technical barriers remain for some time, this avenue cannot be ruled out and should be cautiously observed. If configurations of

AI decision-making with nuclear forces were developed, this could introduce considerable new risks of false alarms, or of *accidental* escalation—especially given the history of cascading 'normal accidents' that have affected nuclear forces.[182]

Human decision-making under pressure

More broadly, the inclusion of AI technology in NC3 may increase the pace of conflicts, reducing the time frame in which decisions can occur and increasing the potential likelihood for inadvertent or accidental escalation.[183] As the perception of an adversary's capabilities are equally as important in deterrence efforts as their actual capabilities, a military's understanding of what (their or their adversaries') military AI systems are in fact able to accomplish may also spur miscalculation and inadvertent escalation.[184] Therefore, AI systems might not need to be deployed to create a destabilising nuclear scenario, as long as they are perceived as creating additional pressures that can lead to miscalculation, or rushed and ill-informed actions.[185]

AI in systems peripheral to NC3

Furthermore, AI does not need to be directly integrated into NC3 itself in order to affect the risks of nuclear war. As noted by Avin and Amadae, while there has been extensive attention on first-order effects of introducing technologies into nuclear command-and-control and weapon-delivery systems, there are also higher-order effects which "stem from the introduction of such technologies into more peripheral systems, with an indirect (but no less real) effect on nuclear risk".[186] For instance, even if militaries believe that AI is not usable for direct nuclear targeting or command, AI systems can still bring about cascading effects through their integration into systems that peripherally impact the safe and secure functioning of NC3; these might include electrical grids, computer systems providing access to relevant intelligence, or weapon platforms associated with the transportation, delivery, or safekeeping of nuclear warheads.

AI as threat to the information environment and accurate intelligence

A fourth avenue of risk is regarding AI's effects on the broader information environment surrounding, framing, and informing nuclear decision-making. In recent years, researchers have begun to explore the ways in which novel AI tools can enable disinformation,[187] and how this may affect societies' epistemic security[188] in ways that make it harder to agree on truth and take coordinated actions that could be crucial for societies to mitigate GCRs (whether this includes coordinated de-escalation around nuclear risks, or other coordination to mitigate other GCRs). For instance, Favaro has mapped how a range of technologies, including AI, might serve as Weapons of Mass Distortion.[189] She distinguishes four clusters of technological effects on the information environment—those that "distort", "compress", "thwart", or "illuminate". A more contested or unclear information environment would also open up new attack surfaces that could be exploited by third-party actors to trigger catalytic escalation amongst its adversaries.

AI as cyber threat to NC3 integrity

Whereas some AI uses within NC3 might be dangerous because of the vulnerabilities they create (as failure points, human decision compressors, or attack surfaces), another channel could involve the use of AI as a *tool* for attacking NC3 systems (regardless of whether they involve AI). This could involve the use of AI-enabled cyber capabilities to attack and disrupt NC3.[190] Experts are increasingly concerned that NC3 is vulnerable to cyberattacks, and that the resulting escalation or unauthorised launch could potentially trigger a GCR scenario.[191] AI technology has been shown to be capable of facilitating increasingly powerful and sophisticated cyberattacks, with increased precision, scope, and scale.[192] Although there is no evidence of states systematically deploying AI-enabled cyber-offensive weapons to date, the convergence of AI and cyber-offensive tools could exacerbate the vulnerabilities of NC3.[193] This could lead to deliberate escalation of offensive cyber-security strategies.[194]

Cyber attacks also can be hard to detect and attribute (quickly);[195] therefore they may be misconstrued, leading to unintentional or catalytic escalation. For example, an offensive operation targeting dual-use

conventional assets could be interpreted as an attack on NC3.[196] It is also broadly agreed that AI acts as a force multiplier for cyber-offensive capabilities.[197] However, it is less clear whether AI will strengthen cyber defence to the same degree as it might strengthen offensive capabilities. The precise effect on the offence-defence balance may be critical to the overall picture.[198] Stronger offensive capabilities could further increase the risk of pre-emptive cyber attacks and subsequently intentional escalation, which would be especially dangerous in the context of nuclear weapon systems.

Broader impacts of AI on nuclear strategic stability

Moreover, the broader deployment of military AI in many other areas could indirectly lead to the disruption of nuclear strategic stability, which could increase the risk of potential intentional or inadvertent escalation.

AI technology could be used to improve a state's capabilities in locating and monitoring an adversary's nuclear second-strike capabilities. For example, better and cheaper autonomous naval drones could track nuclear-armed submarines. This, in turn, could increase the state's perception of likely success in destroying said capabilities before the state's adversary is able to utilise them, and therefore may make a pre-emptive nuclear strike a more attractive strategy than before.[199] Other risks could come from the integration of AI in novel autonomous platforms that are able to operate and loiter in sensitive areas for longer.[200] Even if they were only deployed in order to monitor rival nuclear forces, their pre-positioned presence close to those nuclear assets might prove destabilising, by convincing a defender that they are being deployed to 'scout out' or engage nuclear weapons in advance of a first strike. In these ways, autonomous systems could increase the risks of intentional escalation (when they give a genuine first-strike advantage to one state, or are perceived to do so by another), inadvertent escalation (when errors in their information streams lead to a misinformed decision to launch), or accidental escalation risks, starting the chain of escalation towards a nuclear GCR. Zwetsloot and Dafoe concur that this increased perception of insecurity in nuclear systems could lead to states feeling pressured during times of unrest to engage in pre-emptive escalations.[201]

Finally, in an effort to gain a real or perceived nuclear strategic advantage against their adversaries, while engaging in an AI race, states may place less value on AI safety concerns and more on technological development.[202] This could result in what Danzig has called a "technology roulette"[203] dynamic, with increased risk of prematurely adopting unsafe AI technology in ways that could have profound impacts on the safety or stability of states' nuclear systems.

Contributing factors to AI-nuclear risks

It is important to keep in mind that the risks generated jointly by AI and nuclear weapons are a function of several factors. Firstly, nuclear force posture differs by country, with some forces being more aggressively postured, in ways that enable swifter or immediate use. Additionally, depending on NC3 system design and the degree of force modernisation, AI will interact differently with NC3's component parts—and even dangerously, with brittle legacy systems. Third, the relative robustness or vulnerability of NC3 systems to cyberattacks, for example, will impact systems' resilience to malicious attacks. Along those lines, states' perception of their own vulnerability (as well as the aggressiveness of attackers) will impact stability. This is especially true given that, within complex systems and even through the use of extensive red teaming, it is impossible to identify all system flaws. Fourth, governments' willingness to prematurely deploy AI, either within NC3 and surrounding systems or to augment offensive options for targeting NC3, will be a determinant of catastrophic risk. Fifth, open dialogue, arms control, and risk reduction measures can reduce the potential for nuclear escalation, and a lack of such dialogue can be detrimental. Lastly, luck and normal accidents will inevitably play a role—a fact which highlights unpredictable outcomes amid increased complexity.

Questions for the GCR community

The above discussion has covered a wide range of themes and risk vectors to explore whether—or in what ways—military AI technology is a GCR. Given this, what are the lessons and insights? What policies will be needed to mitigate the potential global catastrophic risks from military AI technology, especially at the intersection with nuclear risk?

Finally, going forward as a field, what are the new lines of research that are needed?

There are lessons specific for the different communities, future questions they should take on, and outlines for an integrated research agenda into military technology, actors, and GCRs that will need further urgent exploration. This chapter has highlighted the urgent need for greater conversation between the different communities engaged on GCRs; on the ethics, safety, and implications of AI; and on nuclear weapons and their risks. We require cross-pollination between these fields, as well as contributions from people with robust expertise in AI and nuclear policy.

In the first place, scholars in defence should reckon with safety and reliability risks around military AI in particular (especially insofar as it poses a GCR), including topics such as robustness, explainability, or susceptibility to adversarial input ('spoofing'). To mitigate these risks, there is value in working with defence industry stakeholders to draw red lines, and to clarify procurement processes.[204]

For nuclear thinkers, there should be greater understanding of the complexities and risks of introducing AI technologies in nuclear weapons. Practically, it will be critical to study how the changing risks of nuclear war—as mediated by AI and machine learning—will impact not just GCR risk, but also the established taboo on nuclear weapons use. How will these changing risks impact governments' calculus about maintaining nuclear arsenals? Are there grounds for optimism about whether or how the 'nuclear taboo' might be elaborated or even extended to a nuclear-AI taboo?

Finally, for experts in both the military and AI fields, more attention needs to be dedicated to investigating the complex and quickly evolving environment that is military AI—especially risks arising at the intersection between nuclear weapons and AI. As made clear in this chapter, concerns around this are not as clear-cut as one might believe upon first glance. Instead, there are a number of possible risk vectors arising from the use of AI throughout the wider landscape, all of which could lead to different forms of nuclear escalation.

In addition, while our analysis in this chapter has made it clear that, at present, there is a small risk of LAWS becoming GCRs, this may not always be the case. It would be useful not only to continue to monitor the development of LAWS to assess if the likelihood of them leading

to global catastrophic events alters, but also to find out how they may interact with other potential GCRs. For example, what might be the possibility of using LAWS to deliver WMDs, and what kind of risk impact could the combination of the two feasibly have? This is another potentially worthwhile avenue for future research.

It is clear that the question 'Is military AI a GCR?' is not only complicated to address, but also a moving target owing to the rapidly evolving technology and risk landscape. To be clear: our preliminary analysis in this chapter has suggested that not all military AI applications qualify as GCRs; however, it also highlights that there are distinct pathways of concern. This is especially the case where emerging military AI technologies intersect with the existing arsenals and command infrastructures of established GCR-level technologies—most notably nuclear weapons. All in all, we invite scholars and practitioners from across the defence studies, GCR, and AI fields (and beyond) to take up the aforementioned challenges, ensuring that this next chapter in global technological risk is not the final one.

Acknowledgements

The authors thank the editors, and in particular SJ Beard and Clarissa Rios Rojas, for their feedback and guidance. For additional and particularly detailed comments on earlier drafts of this chapter, we thank Haydn Belfield and Eva Siegmaim. Matthijs Maas also thanks Seth Baum and Uliana Certan (Global Catastrophic Risk Institute) for conversations and parallel work that clarified some of the arguments and shape of this debate.

Notes and References

1 Beard, S.J. and R. Bronson, 'The story so far: how humanity avoided existential catastrophe', in this volume.

2 Picker, C.B., 'A view from 40,000 feet: International law and the invisible hand of technology', *Cardozo Law Rev., 23* (2001), pp.151–219; Allenby, B., 'Are new technologies undermining the laws of war?', *Bull. At. Sci., 70* (2014), pp.21–31; Deudney, D., 'Turbo change: Accelerating technological

disruption, planetary geopolitics, and architectonic metaphors', *Int. Stud. Rev., 20* (2018), pp.223–31.

3 Boulanin, V. and M. Verbruggen, *Mapping the Development of Autonomy in Weapon Systems, 147* (2017). https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf. Crootof, R., 'The killer robots are here: Legal and policy implications', *CARDOZO LAW Rev., 36* (2015), p.80; Haner, J. and D. Garcia, 'The artificial Intelligence arms race: Trends and world leaders in autonomous weapons development', *Glob. Policy, 10* (2019), pp.331–37.

4 De Spiegeleire, S., M.M. Maas, and T. Sweijs, *Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-Sized Force Providers*. The Hague Centre for Strategic Studies (2017).

5 Amodei, D. et al., *Concrete Problems in AI Safety* (2016). Lehman, J. et al., 'The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities', *Artif. Life, 26* (2019); Arthur Holland, M., *Known Unknowns: Data Issues and Military Autonomous Systems*. (2021). https://www.unidir.org/known-unknowns

6 Belfield, H., A. Jayanti, and S. Avin, *Written Evidence to the UK Parliament Defence Committee's Inquiry on Defence Industrial Policy: Procurement and Prosperity* (2020). https://committees.parliament.uk/writtenevidence/4785/default/

7 Horowitz, M.C. and L. Kahn, 'How Joe Biden can use confidence-building measures for military uses of AI', *Bull. At. Sci., 77* (2021), pp.33–35. Horowitz, M.C. and P. Scharre, *AI and International Stability: Risks and Confidence-Building Measures* (2021). https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures

8 Yudkowsky, E., 'Artificial intelligence as a positive and negative factor in global risk', *Global Catastrophic Risks*. Oxford University Press (2008), pp.308–45; Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (2014); Ngo (2020); Russell (2019); Burden, Clarke, & Whittlestone (2022).

9 Goertzel, B., 'Artificial general intelligence: Concept, state of the art, and future prospects', *J. Artif. Gen. Intell,. 5* (2014), pp.1–48.

10 Grace, K., J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, 'When will AI exceed human performance? Evidence from AI experts', *J. Artif. Intell. Res., 62* (2018), pp.729–54.

11 Gruetzemacher, R. and J. Whittlestone, 'The transformative potential of artificial intelligence', *Futures, 135* (2022), p.102884.

12 Russell, S., *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking (2019); Burden, J., S. Clarke, and J, 'From Turing's speculations to an academic discipline: A history of AI existential safety', in this volume.

13 Future of Life Institute, 'AI safety myths', *Future of Life Institute* (2016). https://futureoflife.org/background/aimyths/

14 Cremer, C.Z., 'Deep limitations? Examining expert disagreement over deep learning', *Prog. Artif. Intell., 10* (2021), pp. 449–64. https://doi.org/10.1007/s13748-021-00239-1

15 Karnofsky, H., 'AI timelines: Where the arguments, and the 'experts,' stand', *Cold Takes* (2021). https://www.cold-takes.com/where-ai-forecasting-stands-today/

16 Hendrycks, D., N. Carlini, J. Schulman, and J. Steinhardt, 'Unsolved problems in ML safety', *ArXiv210913916 Cs* (2021); Amodei (2016).

17 Amodei, D. and J. Clark, 'Faulty reward functions in the wild', *OpenAI* (2016). https://openai.com/blog/faulty-reward-functions/; Krakovna, V. et al., 'Specification gaming: The flip side of AI ingenuity', *Deepmind* (2020). https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity; Turner, A.M., L. Smith, R. Shah, A. Critch, and P. Tadepalli, 'Optimal policies tend to seek power', *ArXiv191201683 Cs* (2021).

18 Cotra, A., 'Why AI alignment could be hard with modern deep learning', *Cold Takes* (2021). https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning

19 Ngo, R., *AGI Safety From First Principles*. (2020).

20 Turchin, A., *Could Slaughterbots Wipe Out Humanity? Assessment of the Global Catastrophic Risk Posed by Autonomous Weapons* (2018); Turchin, A. and D. Denkenberger, 'Military AI as a convergent goal of self-improving AI', in R. Yampolskiy, *Artificial Intelligence Safety and Security*. CRC Press (2018); Vold, K. and D.R. Harris, 'How does artificial intelligence pose an existential risk?', in C. Veliz, *Oxford Handbook of Digital Ethics*. Oxford University Press (2021).

21 Levin, J.-C. and M.M. Maas, 'Roadmap to a roadmap: How could we tell when AGI is a 'Manhattan project' away?', 1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020. (2020).

22 Grace, K., *Leó Szilárd and the Danger of Nuclear Weapons: A Case Study in Risk Mitigation* (2015). https://intelligence.org/files/SzilardNuclearWeapons.pdf

23 Maas, M.M., 'How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons', *Contemp. Secur. Policy, 40* (2019), pp.285–311; Zaidi, W. and A. Dafoe, *International Control of Powerful Technology: Lessons from the Baruch Plan* (2021). https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf

24 Leung, J., *Who Will Govern Artificial Intelligence? Learning From the History of Strategic Politics in Emerging Technologies*. University of Oxford (2019). Ding, J. and A. Dafoe, 'The logic of strategic assets: From oil to AI', *Secur. Stud.* (2021), pp.1–31. https://doi.org/10.1080/09636412.2021.1915583; Ding, J. and A. Dafoe, 'Engines of power: Electricity, AI, and general-purpose military transformations', *ArXiv210604338 Econ Q-Fin* (2021).

25 Beard & Bronson (2023).

26 Baum, S.D. and A.M. Barrett, 'Global catastrophes: The most extreme risks', in V. Bier, *Risks in Extreme Environments: Preparing, Avoiding, Mitigating, and Managing*. Routledge (2018), pp.174–84.

27 Bostrom, N. and M.M. Ćirković, 'Introduction', *Global Catastrophic Risks.* Oxford University Press (2011).

28 Weinberger, S. *The Imagineers of War: The Untold Story of DARPA, the Pentagon Agency That Changed the World*. Random House LLC (2017); Roland, A. and P. Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*. MIT Press (2002).

29 Gordon, T.J. and O. Helmer, *Report on a Long-Range Forecasting Study* (1964). http://stat.haifa.ac.il/~gweiss/courses/OR-logistics/Rand.pdf

30 Defense Science Board Task Force, *Report of the Defense Science Board Task Force on Command and Control Systems Management, 49* (1978).

31 Roland & Shiman (2002).

32 Biddle, S., 'Victory misunderstood: What the Gulf War tells us about the future of conflict', *Int. Secur., 21* (1996), pp.139–79.

33 Cross, S.E. and E. Walker, 'DART: Applying knowledge based planning and scheduling to CRISIS action planning', in M. Zweben and M. Fox, *Intelligent Scheduling*. Morgan Kaufmann (1994), pp.711–29; Hedberg, S.R., 'DART: Revolutionizing logistics planning', *IEEE Intell. Syst., 17* (2002), pp.81–83.

34 Scharre, P., *Autonomous Weapons and Operational Risk* (2016). https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf

35 Scharre, P., *Autonomous Weapons and Stability*. King's College (2020).

36 Scharre (2020).

37 Haner & Garcia (2019).

38 Research and Markets Ltd., *Artificial Intelligence in Military Market by Offering (Software, Hardware, Services), Technology (Machine Learning, Computer vision), Application, Installation Type, Platform, Region—Global Forecast to 2025* (2021). https://www.researchandmarkets.com/reports/5306656/artificial-intelligence-in-military-market-by

39 Haner & Garcia (2019).

40 Haner, J.K., *Dark Horses in the Lethal AI Arms Race* (2019). https://justinkhaner.com/aiarmsrace

41 Trajtenberg, M., *AI as the Next GPT: A Political-Economy Perspective* (2018). http://www.nber.org/papers/w24245 doi:10.3386/w24245; Leung (2019)

42 Maas, M.M., *Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks*. University of Copenhagen (2020); Drezner, D.W., 'Technological change and international relations', *Int. Relat., 33* (2019), pp.286–303.

43 Morgan, F.E. et al., *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World, 224* (2020). https://www.rand.org/content/dam/rand/pubs/research_reports/RR3100/RR3139-1/RAND_RR3139-1.pdf

44 Allen, G. and T. Chan, *Artificial Intelligence and National Security* (2017). http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf

45 Kania, E., 'AlphaGo and beyond: The Chinese military looks to future "intelligentized" warfare', *Lawfare* (2017). https://www.lawfareblog.com/alphago-and-beyond-chinese-military-looks-future-intelligentized-warfare

46 Ding & Dafoe (2021).

47 Nelson, A.J., *The Impact of Emerging Technologies on Arms Control Regimes* (2018).

48 Allen & Chan, (2017).

49 Horowitz, M.C., 'Do emerging military technologies matter for international politics?', *Annu. Rev. Polit. Sci., 23* (2020), pp.385–400.

50 Defense Science Board, *Defense Science Board Summer Study on Autonomy* (2016). https://www.hsdl.org/?abstract&did=794641

51 Soare, S.R., *Digital Divide? Transatlantic Defence Cooperation on Artificial Intelligence* (2020). https://www.iss.europa.eu/content/digital-divide-transatlantic-defence-cooperation-ai

52 Adamsky, D., *The Culture of Military Innovation: The Impact of Cultural Factors on the Revolution in Military Affairs in Russia, the US, and Israel.* Stanford University Press (2010).

53 Verbruggen, M., 'The role of civilian innovation in the development of lethal autonomous weapon systems', *Glob. Policy, 10* (2019), pp.338–42.

54 Soare (2020).

55 Gilli, A., *Preparing for "NATO-mation": The Atlantic Alliance Toward the Age of Artificial Intelligence* (2019). http://www.ndc.nato.int/news/news.php?icode=1270

56 Verbruggen, M., 'Drone swarms: Coming (sometime) to a war near you. Just not today', *Bulletin of the Atomic Scientists* (2021). https://thebulletin.org/2021/02/drone-swarms-coming-sometime-to-a-war-near-you-just-not-today/

57 Gilli, A. and M. Gilli, 'Why China has not caught up yet: military-technological superiority and the limits of imitation, reverse engineering, and cyber espionage', *Int. Secur., 43* (2019), pp.141–89.

58 Amodei, D. and D. Hernandez, 'AI and compute', *OpenAI Blog* (2018). https://openai.com/research/ai-and-compute

59 Ayoub, K. and K. Payne, 'Strategy in the age of artificial intelligence', *J. Strateg. Stud., 39* (2016), pp.793–819; Horowitz, M.C., 'Artificial intelligence, international competition, and the balance of power', *Texas National Security Review* (2018).

60 Dafoe, A., 'On technological determinism: A typology, scope conditions, and a mechanism', *Sci. Technol. Hum. Values, 40* (2015), pp.1047–076.

61 Maas (2019).

62 Bleek, P.C., *When Did (and Didn't) States Proliferate? Chronicling the Spread of Nuclear Weapons, 56* (2017). https://www.belfercenter.org/sites/default/files/files/publication/When%20Did%20%28and%20Didn%27t%29%20States%20Proliferate%3F_1.pdf; Meyer, S., S. Bidgood, and W.C. Potter, 'Death dust: The little-known story of US and Soviet pursuit of radiological weapons', *Int. Secur., 45* (2020), pp.51–94.

63 Rosert, E. and F. Sauer, 'How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies', *Contemp. Secur. Policy, 0* (2020), pp.1–25. https://www.tandfonline.com/doi/full/10.1080/13523260.2020.1771508; Belfield, H., 'Activism by the AI community:

Analysing recent achievements and future prospects', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM (2020), pp.15–21. https://doi.org/10.1145/3375627.3375814

64 McDonald, J., *What If Military AI Is a Washout?* (2021). https://jackmcdonald.org/book/2021/06/what-if-military-ai-sucks/

65 McDonald (2021).

66 Garfinkel, B. and A. Dafoe, 'How does the offense-defense balance scale?', *J. Strateg. Stud., 42* (2019), pp.736–63; Lieber, K.A., 'Grasping the technological peace: the offense-defense balance and international security', *Int. Secur., 25* (2000), p.71.

67 Payne, K., *I, Warbot: The Dawn of Artificially Intelligent Conflict*. C Hurst & Co Publishers Ltd (2021).

68 Boulanin & Verbruggen (2017).

69 Blain, L., *South Korea's Autonomous Robot Gun Turrets: Deadly From Kilometers Away* (2010). http://newatlas.com/korea-dodamm-super-aegis-autonomos-robot-gun-turret/17198/; Velez-Green, A., 'The foreign policy essay: The South Korean sentry—a "killer robot" to prevent war', *Lawfare* (2015). https://www.lawfareblog.com/foreign-policy-essay-south-korean-sentry%E2%80%94-killer-robot-prevent-war

70 Israel Aerospace Industries, *Harpy Loitering Weapon* (2020). https://www.iai.co.il/p/harpy.

71 Future of Life Institute, '5 real-life technologies that prove autonomous weapons are already here', *Future of Life Institute* (2021). https://futureoflife.org/2021/11/22/5-real-life-technologies-that-prove-autonomous-weapons-are-already-here/

72 Tucker, P., 'SecDef: China is exporting killer robots to the Mideast', *Defense One* (2019). https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/

73 Trevithick, J., 'Turkey now has swarming suicide drones it could export', *The Drive* (2020).

74 UN Panel of Experts on Libya, *Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973* (2021). https://undocs.org/pdf?symbol=en/S/2021/229; Cramer, M., 'A.I. Drone may have acted on its own in attacking fighters, U.N. says', *The New York Times* (2021).

75 Kesteloo, H., 'Punisher drones are positively game-changing for Ukrainian military in fight against Russia', *DroneXL* (2022). https://dronexl.co/2022/03/03/punisher-drones-ukrainian-military/; Trabucco, L. and

K.J. Heller, 'Beyond the ban: Comparing the ability of 'killer robots' and human soldiers to comply with IHL', *Fletcher Forum World Aff. 46* (2022).

76 Scharre, P., *Robotics on the Battlefield, Part II: The Coming Swarm, 68* (2014). https://www.cnas.org/publications/reports/robotics-on-the-battlefield-part-ii-the-coming-swarm

77 Hambling, D., 'Israel's combat-proven drone swarm may be start of a new kind of warfare', *Forbes* (2021). https://www.forbes.com/sites/davidhambling/2021/07/21/israels-combat-proven-drone-swarm-is-more-than-just-a-drone-swarm/

78 Michel, A.H., 'The killer algorithms nobody's talking about', *Foreign Policy* (2020). https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/

79 Bergman, R. and F. Fassihi, 'The scientist and the A.I.-assisted, remote-control killing machine', *The New York Times* (2021).

80 Knight, W., 'A dogfight renews concerns about AI's lethal potential', *Wired* (2020).

81 US Department of Defense, *Department of Defense Announces Successful Micro-Drone Demonstration.* US Department of Defense (2017). https://www.defense.gov/News/Releases/Release/Article/1044811/department-of-defense-announces-successful-micro-drone-demonstration/

82 Chavannes, E., K. Klonowska, and T. Sweijs, 'Governing autonomous weapon systems: Expanding the solution space, from scoping to applying', *HCSS Secur., 39* (2020); Rosert, & Sauer (2020).

83 Carpenter, C., *'Lost' Causes, Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Cornell University Press (2014). https://doi.org/10.7591/9780801470363

84 Liu, H.-Y., 'Categorization and legality of autonomous and remote weapons systems', *Int. Rev. Red Cross, 94* (2012), pp.627–52; Anderson, K., D. Reisner, and M. Waxman, 'Adapting the law of armed conflict to autonomous weapon systems', *Int. Law Stud., 90* (2014), p.27.

85 Human Rights Watch, *Shaking the Foundations: The Human Rights Implications of Killer Robots* (2014).

86 Rosert, E. and F. Sauer, 'Prohibiting autonomous weapons: Put human dignity first', *Glob. Policy, 10* (2019), pp.370–75; Rosert, & Sauer (2020).

87 Kallenborn, Z., 'Meet the future weapon of mass destruction, the drone swarm', *Bulletin of the Atomic Scientists* (2021). https://thebulletin.org/2021/04/meet-the-future-weapon-of-mass-destruction-the-drone-swarm/

88 Bahçecik, Ş.O., 'Civil society responds to the AWS: Growing activist networks and shifting frames, *Glob. Policy,* 10(3) (2019), pp. 365–69.

89 Future of Life Institute, *Slaughterbots* (2017). Future of Life Institute, *Slaughterbots—If human: kill*() (2021).

90 Bahçecik (2019).

91 Turchin, A. and D. Denkenberger, 'Classification of global catastrophic risks connected with artificial intelligence', *AI Soc., 35* (2020), pp.147–63.

92 Kallenborn, Z. and P.C. Bleek, 'Swarming destruction: drone swarms and chemical, biological, radiological, and nuclear weapons', *Nonproliferation Rev., 25* (2018), pp.523–43.

93 Kunz, M. and S. Ó hÉigeartaigh, 'Artificial intelligence and robotization', in R. Geiss and N. Melzer, *Oxford Handbook on the International Law of Global Security*. Oxford University Press (2021).

94 Rogers, J., 'The dark side of our drone future', *Bulletin of the Atomic Scientists* (2019). https://thebulletin.org/2019/10/the-dark-side-of-our-drone-future/

95 Mani, L., A. Tzachor, and P. Cole, 'Global catastrophic risk from lower magnitude volcanic eruptions', *Nat. Commun., 12* (2021), p.4756.

96 Solodov, A., A. Williams, S.A. Hanaei, and B. Goddard, 'Analyzing the threat of unmanned aerial vehicles (UAV) to nuclear facilities', *Secur. J. Lond. 31* (2018), pp.305–24.

97 Tang, A. and L. Kemp, 'A fate worse than warming? Stratospheric aerosol injection and global catastrophic risk', *Front. Clim., 3* (2021), p.144; Baum, S.D., T.M. Maher, and J. Haqq-Misra, 'Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse', *Environ. Syst. Decis. 33* (2013), pp.168–80.

98 Future of Life Institute, *An Open Letter to the United Nations Convention on Certain Conventional Weapons* (2017). https://futureoflife.org/autonomous-weapons-open-letter-2017/

99 Russell, S., A. Aguirre, A. Conn, and M. Tegmark, 'Why you should fear "slaughterbots"—a response', *IEEE Spectr.* (2018).

100 Turchin, &. Denkenberger (2020).

101 Aguirre, A., 'Why those who care about catastrophic and existential risk should care about autonomous weapons', *EA Forum* (2020). https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2

102 Turchin (2018).

103 Ibid.

104 Aguirre (2020).

105 Brookings, *What Nuclear Weapons Delivery Systems Really Cost* (2016). https://www.brookings.edu/what-nuclear-weapons-delivery-systems-really-cost/

106 Blumberg, Y., 'Here's how much a nuclear weapon costs', *CNBC* (2017). https://www.cnbc.com/2017/08/08/heres-how-much-a-nuclear-weapon-costs.html

107 UN Secretary-General, *Chemical and Bacteriological* (*Biological*) *Weapons and the Effects of Their Possible Use* (1969). Koblentz, G.D., *Living Weapons: Biological Warfare and International Security*. Cornell University Press (2011).

108 Scharre (2014).

109 Baum, S.D., A.M. Barrett, U. Certan, and M.M. Maas, *Autonomous Weapons and the Long-Term Future* (2022).

110 Sabbagh, D., 'Killer drones: How many are there and who do they target?', *The Guardian* (2019).

111 Vailshery, L.S., 'Global consumer drone shipments 2020–2030', *Statista* (2021). https://www.statista.com/statistics/1234658/worldwide-consumer-drone-unit-shipments

112 Hambling, D., 'The U.S. Navy plans to foil massive 'super swarm' drone attacks by using the swarm's intelligence against itself', *Forbes* (2020).

113 Turchin (2018).

114 Kemp, L., 'Agents of doom: Who is creating the apocalypse and why', *BBC Future* (2021).

115 Kallenborn (2021).

116 Baum, Barrett, Certan, & Maas, (2022).

117 Perkovich, G., 'Will you listen? A dialogue on creating the conditions for nuclear disarmament', *Carnegie Endowment for International Peace* (2018). https://carnegieendowment.org/2018/11/02/will-you-listen-dialogue-on-creating-conditions-for-nuclear-disarmament-pub-77614

118 Lewis, J., 'Point and nuke: Remembering the era of portable atomic bombs', *Foreign Policy* (2018). https://foreignpolicy.com/2018/09/12/point-and-nuke-davy-crockett-military-history-nuclear-weapons/

119 Galison, P L. and B. Bernstein, 'In any light: Scientists and the decision to build the superbomb, 1952–1954', *Hist. Stud. Phys. Biol.*

Sci., 19 (1989), pp.267–347; Wellerstein, A., 'The leak that brought the H-bomb debate out of the cold', *Restricted Data: The Nuclear Secrecy Blog* (2021). http://blog.nuclearsecrecy.com/2021/06/14/the-leak-that-brought-the-h-bomb-debate-out-of-the-cold/

120 Horgan, J., 'Bethe, Teller, Trinity and the end of Earth', *Scientific American Blog Network* (2015). https://blogs.scientificamerican.com/cross-check/bethe-teller-trinity-and-the-end-of-earth/; Ellsberg, D., *The Doomsday Machine: Confessions of a Nuclear War Planner*. Bloomsbury USA (2017).

121 Scarry, E., *Thermonuclear Monarchy: Choosing Between Democracy and Doom*. W. W. Norton & Company (2016).

122 BBC, 'Hiroshima and Nagasaki: 75th anniversary of atomic bombings', *BBC News* (2020).

123 PBS News Hour, 'Types of nuclear bombs', *PBS NewsHour* (2005). https://www.pbs.org/newshour/nation/military-jan-june05-bombs_05-02

124 Toon, O.B. et al, 'Rapidly expanding nuclear arsenals in Pakistan and India portend regional and global catastrophe', *Sci. Adv., 5* (2019).

125 Rosenberg, D.A. and W.B. Moore, ''Smoking radiating ruin at the end of two hours': Documents on American plans for nuclear war with the Soviet Union, 1954–55', *Int. Secur., 6* (1981), pp.3–38.

126 Rosenberg, D., 'Constraining overkill: Contending approaches to nuclear strategy, 1955–1965', *Naval History and Heritage Command* (1994).

127 Ellsberg (2017).

128 Badash, L., *A Nuclear Winter's Tale: Science and Politics in the 1980s*. MIT Press (2009). Sagan, C., 'Nuclear war and climatic catastrophe: Some policy implications', *Foreign Aff. 62* (1983), pp.257–92.

129 US Secretary of Defense, 'Nuclear winter: The view from the US Defense Department', *Survival, 27* (1985), pp.130–34.

130 Badash (2009).

131 Tannenwald, N., 'The nuclear taboo: The United States and the normative basis of nuclear non-use', *Int. Organ., 53* (1999), pp.433–68.

132 Sauer, F., *Atomic Anxiety: Deterrence, Taboo and the Non-Use of US Nuclear Weapons*. Springer (2015).

133 Robock, A. and O.B. Toon, 'Local nuclear war, global suffering', *Sci. Am., 302* (2010), pp.74–81.

134 Helfand, I., *Nuclear Famine: Two Billion People At Risk? Global Impacts of Limited Nuclear War on Agriculture, Food Supplies, and Human Nutrition*

(2013). https://www.psr.org/wp-content/uploads/2018/04/two-billion-at-risk.pdf

135 Coupe, J., C.G. Bardeen, A. Robock, and O.B. Toon, 'Nuclear winter responses to nuclear war between the United States and Russia in the whole atmosphere community climate model version 4 and the Goddard Institute for Space Studies Model', *Geophys. Res. Atmospheres, 124* (2019), pp.8522–543.

136 Reisner, J. et al., 'Climate impact of a regional nuclear weapons exchange: An improved assessment based on detailed source calculations', *J. Geophys. Res. Atmospheres, 123* (2018), pp.2752–772.

137 Frankel, M., J. Scouras, and G. Ullrich, *The Uncertain Consequences of Nuclear Weapons Use* (2015). https://apps.dtic.mil/sti/citations/ADA618999

138 Scouras, J., 'Nuclear war as a global catastrophic risk', *J. Benefit-Cost Anal., 10* (2019), pp.274–95.

139 Gavin, F.J., 'We need to talk: The past, present, and future of US nuclear weapons policy', *War on the Rocks* (2017). https://warontherocks.com/2017/01/we-need-to-talk-the-past-present-and-future-of-u-s-nuclear-weapons-policy/

140 Rodriguez, L., *How Likely Is a Nuclear Exchange Between the US and Russia?* (2019). https://rethinkpriorities.org/publications/how-likely-is-a-nuclear-exchange-between-the-us-and-russia; Baum, S., R. de Neufville, and A. Barrett, 'A model for the probability of nuclear war', *Glob. Catastrophic Risk Inst. Work. Pap.* (2018). https://doi.org/10.2139/ssrn.3137081; Baum, S., 'Reflections on the risk analysis of nuclear war', in B.J. Garrick, *Proceedings of the Workshop on Quantifying Global Catastrophic Risks*. Garrick Institute for the Risk Sciences (2018), pp.19–50.

141 Schlosser, E., *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. Penguin Books (2014); Sagan, S.D., *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton University Press (1993). https://doi.org/10.1515/9780691213064

142 Sagan, S.D., 'Learning from normal accidents', *Organ. Environ., 17* (2004), pp.15–19.

143 Maas (2019).

144 Maas, M.M., 'Regulating for "normal AI accidents": Operational lessons for the responsible governance of artificial intelligence deployment', *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery (2018), pp.223–28. https://doi.org/10.1145/3278721.3278766

145 Rodriguez, L., 'What is the likelihood that civilizational collapse would directly lead to human extinction (within decades)?', *Effective Altruism Forum* (2020). https://forum.effectivealtruism.org/posts/GsjmufaebreiaivF7/what-is-the-likelihood-that-civilizational-collapse-would; Wiblin, R., *Luisa Rodriguez on Why Global Catastrophes Seem Unlikely to Kill Us All*. 80,000 Hours Podcast (2021).

146 Belfield, H., 'Collapse, recovery and existential risk', in P. Callahan, M. Centeno, P. Larcey, and T. Patterson, *The End of the World as We Know It*. Routledge Press (2022).

147 Schubert, S., L. Caviola, and N.S. Faber, 'The psychology of existential risk: Moral judgments about human extinction', *Sci. Rep., 9* (2019), p.15100.

148 Kohler, S., *Cooperative Security and the Nunn-Lugar Act*. *4* (1989).

149 Wunderlich, C., H. Müller, and U. Jakob, *WMD Compliance and Enforcement in a Changing Global Context* (2020). https://www.unidir.org/publication/wmd-compliance-and-enforcement-changing-global-context

150 Reif, K. and S. Bugos, 'US, Russia extend new START for five years', *Arms Control Association* (2021). https://www.armscontrol.org/act/2021-03/news/us-russia-extend-new-start-five-years

151 Kühn, U., 'Why arms control is (almost) dead', *Carnegie Europe* (2020). https://carnegieeurope.eu/strategiceurope/81209

152 Wan, W., *Nuclear Escalation Strategies and Perceptions: The United States, the Russian Federation, and China* (2021). https://unidir.org/escalation; https://doi.org/10.37559/WMD/21/NRR/02

153 Lucero-Matteucci, K.T., 'Signs of life in nuclear diplomacy: A look beyond the doom and gloom', *Georget. J. Int. Aff* (2019).

154 Kirstensen, H.M., 'US deploys new low-yield nuclear submarine warhead', *Federation Of American Scientists* (2020). https://fas.org/blogs/security/2020/01/w76-2deployed/

155 Fink, A.L. and O. Oliker, 'Russia's nuclear weapons in a multipolar world: guarantors of sovereignty, great power status & more', *Daedalus, 149* (2020), pp.37–55.

156 Piotrowski, M.A., *Russia's Status-6 Nuclear Submarine Drone* (*Poseidon*) (2018) https://pism.pl/publications/Russia_s_Status_6_Nuclear_Submarine_Drone__Poseidon_

157 Vaddi, P., 'Bringing Russia's new nuclear weapons into new START', *Carnegie Endowment for International Peace* (2019). https://carnegieendowment.org/2019/08/13/bringing-russia-s-new-nuclear-weapons-into-new-start-pub-79672

158 Edmonds, J. et al., *Artificial Intelligence and Autonomy in Russia* 258 (2021). https://www.cna.org/CNA_files/centers/CNA/sppp/rsp/russia-ai/Russia-Artificial-Intelligence-Autonomy-Putin-Military.pdf

159 Kristensen, H.M. and M. Korda, 'Chinese nuclear weapons, 2021', *Bull. At. Sci., 77* (2021), pp.318–36.

160 Kristensen, H.M. and M. Korda, 'China's nuclear missile silo expansion: from minimum deterrence to medium deterrence', *Bulletin of the Atomic Scientists* (2021). https://thebulletin.org/2021/09/chinas-nuclear-missile-silo-expansion-from-minimum-deterrence-to-medium-deterrence/

161 Wright, T., 'Is China gliding toward a FOBS capability?', *IISS* (2021). https://www.iiss.org/blogs/analysis/2021/10/is-china-gliding-toward-a-fobs-capability. Acton, J.M., 'China's tests are no Sputnik moment', *Carnegie Endowment for International Peace* (2021). https://carnegieendowment.org/2021/10/21/china-s-tests-are-no-sputnik-moment-pub-85625

162 Mills, C., *Integrated Review 2021: Increasing the Cap on the UK's Nuclear Stockpile* (2021).

163 Acton, J.M. *Is It a Nuke?: Pre-Launch Ambiguity and Inadvertent Escalation* (2020). https://carnegieendowment.org/2020/04/09/is-it-nuke-pre-launch-ambiguity-and-inadvertent-escalation-pub-81446

164 Futter, A. and B. Zala, 'Strategic non-nuclear weapons and the onset of a Third Nuclear Age', *Eur. J. Int. Secur., 6* (2021), pp.257–77.

165 Trabucco, L. and M.M. Maas, 'Into the thick of it: Mapping the emerging landscape of military AI strategic partnerships', in AutoNorms / Center for War Studies (SDU) conference "The Algorithmic Turn in Security and Warfare" (2022); Stanley-Lockman, Z., *Military AI Cooperation Toolbox: Modernizing Defense Science and Technology Partnerships for the Digital Age* (2021). https://cset.georgetown.edu/wp-content/uploads/CSET-Military-AI-Cooperation-Toolbox.pdf; Bendett, S. and E.B. Kania, *A new Sino-Russian High-Tech Partnership: Authoritarian Innovation in an Era of Great-Power Rivalry, 24* (2019). https://www.aspi.org.au/report/new-sino-russian-high-tech-partnership

166 Thompson, N. and I. Bremmer, 'The AI Cold War that threatens us all', *Wired* (2018).

167 Borning, A., 'Computer system reliability and nuclear war', *Commun. ACM, 30* (1987), pp.112–31.

168 Raushenbakh, B.V., 'Computer war', in A.A. Gromyko and M. Hellman (eds), *Breakthrough: Emerging New Thinking: Soviet and Western scholars Issue a Challenge to Build a World Beyond War*. Walker (1988).

169 Hoffman, D., *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. Anchor (2010).

170 Harvey, J.R., 'US nuclear command and control for the 21st century', *Nautilus Institute for Security and Sustainability* (2019). https://nautilus.org/napsnet/napsnet-special-reports/u-s-nuclear-command-and-control-for-the-21st-century/

171 Cunningham, F., 'Nuclear command, control, and communications systems of the People's Republic of China', *Nautilus Institute for Security and Sustainability* (2019). https://nautilus.org/napsnet/napsnet-special-reports/nuclear-command-control-and-communications-systems-of-the-peoples-republic-of-china/

172 Gower, J., *United Kingdom: Nuclear Weapon Command, Control, and Communications* (2019). https://securityandtechnology.org/wp-content/uploads/2020/07/gower_uk_nc3_report_IST.pdf

173 Johnson, J., 'Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age', *Eur. J. Int. Secur.* (2021a) pp.1–23. https://doi.org/10.1017/eis.2021.23; Johnson, J., ''Catalytic nuclear war' in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states', *J. Strateg. Stud.*, (2021b), pp.1–41.

174 Johnson (2021a).

175 Johnson (2021b).

176 Geist, E. and A.J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?*, *28* (2018). https://www.rand.org/pubs/perspectives/PE296.html; Fitzpatrick, M., 'Artificial intelligence and nuclear command and control', *Survival, 61* (2019), pp.81–92; Field, M., 'Strangelove redux: US experts propose having AI control nuclear weapons', *Bulletin of the Atomic Scientists* (2019). https://thebulletin.org/2019/08/strangelove-redux-us-experts-propose-having-ai-control-nuclear-weapons/; Johnson, J., 'Delegating strategic decision-making to machines: Dr. Strangelove Redux?', *J. Strateg. Stud.*, 45(3) (2022), pp.439–77.

177 Despite speculation that Russia's 'Dead Hand' system is still in use, there is no definitive evidence that this is or will continue to be the case.

178 Lowther, A. and C. McGiffin, 'America needs a "Dead Hand"', *War on the Rocks* (2019). https://warontherocks.com/2019/08/america-needs-a-dead-hand/

179 Freedberg, S.J., 'No AI for nuclear command & control: JAIC's Shanahan', *Breaking Defense* (2019). https://breakingdefense.sites.breakingmedia.com/2019/09/no-ai-for-nuclear-command-control-jaics-shanahan/

180 Fedasiuk, R., 'We spent a year investigating what the Chinese army is buying. Here's what we learned', *POLITICO* (2021). Fedasiuk, R., J. Melot, and B. Murphy, *Harnessed Lightning: How the Chinese Military is Adopting Artificial Intelligence* (2021). https://cset.georgetown.edu/publication/harnessed-lightning/

181 Loss, R. and J. Johnson, 'Will artificial intelligence imperil nuclear deterrence?', *War on the Rocks* (2019). https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/

182 Maas (2019). Schlosser (2014). Sagan, (1993).

183 Johnson, J.S., 'Artificial intelligence: A threat to strategic stability', *Strateg. Stud. Q.* (Spring 2020), pp.16–39. Payne, K., 'Artificial intelligence: A revolution in strategic affairs?', *Survival, 60* (2018), pp.7–32.

184 Johnson (2020).

185 Amadae, S.M. et al., *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. 1*. SIPRI (2019).

186 Avin, S. and S.M. Amadae, 'Autonomy and machine learning at the interface of nuclear weapons, computers and people', in V. Boulanin (ed), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Stockholm International Peace Research Institute (2019). https://doi.org/10.17863/CAM.44758

187 Citron, D. and R. Chesney, 'Deepfakes and the new disinformation war: The coming age of post-truth geopolitics', *Foreign Affairs, 98* (2019).

188 Epistemic security has been defined as the state which "ensures that a community's processes of knowledge production, acquisition, distribution, and coordination are robust to adversarial (or accidental) influence [such that] [e]pistemically secure environments foster efficient and well-informed group decision-making which helps decision-makers to better achieve their individual and collective goals". Seger, E. et al., *Tackling Threats to Informed Decisionmaking in Democratic Societies: Promoting Epistemic Security in a Technologically-Advanced World* (2020). https://www.turing.ac.uk/research/publications/tackling-threats-informed-decision-making-democratic-societies

189 Favaro, M., *Weapons of Mass Distortion: A New Approach to Emerging Technologies, Risk Reduction, and the Global Nuclear Order* (2021). https://www.kcl.ac.uk/csss/assets/weapons-of-mass-distortion.pdf

190 Johnson, J. and E. Krabill, 'AI, cyberspace, and nuclear weapons', *War on the Rocks* (2020). https://warontherocks.com/2020/01/ai-cyberspace-and-nuclear-weapons/

191 Sharikov, P., 'Artificial intelligence, cyberattack, and nuclear weapons—a dangerous combination', *Bull. At. Sci., 74* (2018), pp.368–73.

192 Schneier, B., *The Coming AI Hackers* (2021). https://www.schneier.com/wp-content/uploads/2021/04/The-Coming-AI-Hackers.pdf; Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (2018). http://arxiv.org/abs/1802.07228

193 Futter, A., *Hacking the Bomb: Cyber Threats and Nuclear Weapons*. Georgetown University Press (2018).

194 Johnson and Krabill (2020).

195 Eilstrup-Sangiovanni, M., 'Why the world needs an international cyberwar convention', *Philos. Technol., 31* (2018), pp.379–407.

196 Johnson, J., 'The AI-cyber nexus: implications for military escalation, deterrence and strategic stability', *J. Cyber Policy* (2019).

197 Gartzke, E. and J.R. Lindsay, 'Weaving tangled webs: Offense, defense, and deception in cyberspace', *Secur. Stud., 24* (2015), pp.316–48.

198 Zwetsloot, R. and A. Dafoe, 'Thinking about risks from AI: Accidents, misuse and structure', *Lawfare* (2019). https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure

199 Geist, E. and A.J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?, 28* (2018). https://www.rand.org/pubs/perspectives/PE296.html . Fitzpatrick (2019).

200 Kallenborn, Z., 'AI risks to nuclear deterrence are real', *War on the Rocks* (2019). https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/

201 Zwetsloot and Dafoe, (2019).

202 Armstrong, S., N. Bostrom, and C. Shulman, 'Racing to the precipice: A model of artificial intelligence development', *AI Soc., 31* (2016), pp.201–06.

203 Danzig, R., *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority, 40* (2018). https://www.cnas.org/publications/reports/technology-roulette

204 Belfield, Jayanti & Avin (2020).